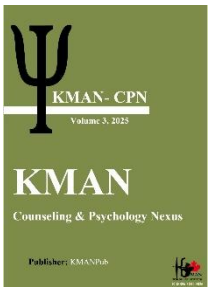
PUBLISHING INSTITUTE
Journal Website

Article history:
Received 02 May 2026
Revised 03 August 2026
Accepted 10 August 2026
Initial Publication 15 August 2026
Final Publication 20 August 2026

KMAN Counseling & Psychology Nexus

OPEN PEER-REVIEW



E-ISSN: 3041-9026

Explainable Semi-Supervised Learning for Depression Subtype Detection in Social Media

Ali. Nakhaei[1], Hamed. Vahdat Nejad[2*], Mahdi. Khazaei Pour[1]

[1] Department of Computer Engineering, Bi.C., Islamic Azad University, Birjand, Iran
[2] Department of Computer Engineering, University of Birjand, Birjand, Iran

*** Corresponding author email address**: vahdatnejad@birjand.ac.ir

Editor	Reviewers
Izet Pehlić Full professor for Educational sciences, Islamic pedagogical faculty of the University of Zenica, Bosnia and Herzegovina izet.pehlic@unze.ba	**Reviewer 1:** Parvaneh Mohammadkhani Professor, Department of Clinical Psychology, University of Rehabilitation Sciences and Social Health, Tehran, Iran. Email: Pa.mohammadkhani@uswr.ac.ir **Reviewer 2:** Farideh Dokanehi Fard Associate Professor, Counseling Department, Roudehen Branch, Islamic Azad University, Roudehen, Iran. Email: f.dokaneifard@riau.ac.ir

1. Round 1

1.1. Reviewer 1

Reviewer:

The Introduction states that the study addresses “five clinically relevant categories: Major Depressive Disorder, Bipolar Depression, Psychotic Depression, Atypical Depression, and Postpartum Depression.” The clinical terminology requires considerably greater precision. These five labels do not occupy equivalent nosological levels: Major Depressive Disorder is a diagnostic category, bipolar depression refers to a depressive episode occurring within bipolar disorder, while psychotic, atypical, and postpartum presentations are generally conceptualized as features/specifiers or clinically defined presentations rather than mutually exclusive standalone depressive disorders. Treating all five as parallel “subtypes” creates a clinically problematic ontology and potentially allows genuine overlap between classes. The manuscript should justify the use of the term “subtype,” describe exactly how Nusrat et al. operationalized each class, discuss whether classes were mutually exclusive by construction, and acknowledge that the computational labels should not be interpreted as equivalent to formal differential psychiatric diagnoses.

The manuscript reports that the labeled corpus was “constructed using subtype-specific lexical patterns curated and validated with psychiatric input.” This characteristic of the dataset is not merely a limitation; it directly affects the construct validity of the entire classification task. If subtype labels were assigned partly through subtype-specific lexical expressions, then a TF-IDF classifier may simply recover the lexical rules used during dataset construction, producing circularity between annotation

and prediction. This concern is particularly important given the very high Macro-F1 of 0.887 even when only 10% of the labeled data were used. The authors should provide a detailed account of how each original label was generated, whether diagnostic keywords used for labeling remained in the input text, whether explicit subtype names or highly deterministic lexical triggers were removed, and whether a sensitivity analysis excluding label-defining terms changes performance. Without such analysis, the reported classification accuracy may substantially overestimate clinically meaningful subtype discrimination.

The same implementation table reports "Training epochs (supervised backbone): 15, with early stopping on validation Macro-F1." This terminology is unusual for a conventional TF-IDF + linear-SVM pipeline because standard batch SVM optimization is not normally reported in epochs with neural-network-style early stopping. The authors should clarify the specific optimization algorithm and software parameters that make "15 epochs" meaningful. If SGDClassifier or another iterative linear optimizer was used, this should replace the generic label "linear SVM," and parameters such as loss function, regularization coefficient, maximum iterations, tolerance, random state, and stopping criterion should be reported. If a conventional SVM implementation was used, the statement about epochs and early stopping appears technically inconsistent and should be corrected.

The preprocessing section states that "All text was lowercased, and URLs, user mentions, and non-informative special characters were removed while preserving emoji and punctuation known to carry affective signal." This description is too general for reproducibility and is particularly consequential for a TF-IDF model whose predictions depend directly on tokenization and surface-form preprocessing. The authors should provide the exact tokenizer, URL and mention replacement strategy, handling of hashtags, contractions, repeated characters, negation, emojis, punctuation, stop words, stemming/lemmatization, Unicode normalization, and minimum/maximum document-frequency thresholds. They should also clarify whether hashtags were removed, segmented, or retained, because subtype-defining hashtags could constitute severe label leakage in a Twitter dataset derived using lexical retrieval criteria.

Response: Revised and uploaded the manuscript.

1.2. Reviewer 2

Reviewer:

In Section 3.2.1, the authors state that "Control-class tweets were intentionally excluded from supervised optimization and instead incorporated into the unlabeled pool." This design decision requires major methodological justification because the semi-supervised classifier has only five depressive output classes. If genuinely non-depressed control observations are entered into an unlabeled pool but the classifier is forced to assign every high-confidence instance to one of five depression subtypes, the self-training procedure can systematically generate false depressive pseudo-labels for control observations. This may provide a direct methodological explanation for the deterioration observed under self-training. The authors should report how many control instances entered each self-training round, how many received pseudo-labels above the confidence threshold, which depressive classes they were assigned to, and whether excluding control observations from the unlabeled pool changes the results. A more principled alternative would be to retain a sixth control class or use an open-set/out-of-distribution rejection mechanism.

Section 4.7 reports that the cleaned corpus contained 14,983 tweets, whereas Section 4.2 initially describes the Nusrat Twitter dataset as containing 23,634 samples. Later, the manuscript explains only that missing or empty entries were removed before arriving at 14,983 observations. A reduction of approximately 8,651 observations is substantial and cannot be adequately described as routine cleaning without a transparent attrition analysis. The authors should provide a reproducible flow diagram or table showing the original number of observations, duplicates removed, missing-text records, non-English records if applicable, posts removed by length restrictions, duplicate/near-duplicate posts, retweets, and any other exclusion criteria. Class-specific attrition should also be reported because differential removal across subtypes could alter class distributions and inflate separability.

E-ISSN: 3041-9026

The manuscript states that "only the five depression subtypes from Nusrat … were used for supervised training" and that the held-out test set represented "20% of the five-subtype data." However, the exact sequence of train/validation/test splitting relative to masking labels and constructing the unlabeled pool is insufficiently specified. This is critical in semi-supervised experiments because leakage can occur if test-set observations—or duplicates/near-duplicates of them—are inadvertently included as unlabeled examples during self-training. The authors should state explicitly whether the test set was permanently isolated before any pseudo-labeling, vectorizer fitting, feature selection, threshold tuning, or self-training iteration. They should also confirm whether TF-IDF vocabulary and inverse-document-frequency statistics were fit exclusively on the training partition rather than on the complete corpus. A precise data-partition schematic is strongly recommended.

Section 4.6 describes the evaluated supervised backbone as "TF-IDF (word 1–2 grams) + linear SVM" and reports a "confidence threshold $\tau = 0.90$." The manuscript does not explain how probabilistic confidence was obtained from the linear SVM. Standard linear SVM decision scores are margins and are not inherently calibrated probabilities bounded in a manner directly interpretable as 0.90 confidence. The authors must specify the exact estimator used, including whether LinearSVC, SVC(kernel='linear', probability=True), SGD-based hinge classification, or another implementation was employed. If probabilities were generated through Platt scaling or another calibration procedure, the calibration set and procedure should be reported. The choice of 0.90 should also be justified empirically, ideally through validation-set sensitivity analyses across multiple confidence thresholds.

Response: Revised and uploaded the manuscript.

2. Revised

Editor's decision after revisions: Accepted.

Editor in Chief's decision: Accepted.

KMAN-CPN
KMAN-Counseling & Psychology Nexus
E-ISSN: 3041-9026